Prevalence of *syn* nucleobases in the active sites of functional RNAs

JOSHUA E. SOKOLOSKI, STEPHANIE A. GODFREY,¹ SARAH E. DOMBROWSKI,² and PHILIP C. BEVILACQUA³

Department of Chemistry and Center for RNA Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

ABSTRACT

Biological RNAs, like their DNA counterparts, contain helical stretches, which have standard Watson-Crick base pairs in the *anti* conformation. Most functional RNAs also adopt geometries with far greater complexity such as bulges, loops, and multihelical junctions. Occasionally, nucleobases in these regions populate the *syn* conformation wherein the base resides close to or over the ribose sugar, which leads to a more compact state. The importance of the *syn* conformation to RNA function is largely unknown. In this study, we analyze 51 RNAs with tertiary structure, including aptamers, riboswitches, ribozymes, and ribosomal RNAs, for number, location, and properties of *syn* nucleobases. These RNAs represent the set of nonoverlapping, moderate- to high-resolution structures available at present. We find that *syn* nucleobases are much more common among purines than pyrimidines, and that they favor C2'-endo-like conformations especially among those nucleobases in the intermediate *syn* conformation. Strikingly, most *syn* nucleobases participate in tertiary stacking and base-pairing interactions: Inspection of RNA structures revealed that the majority of the *syn* nucleobases are in regions assigned to function, with many *syn* nucleobases interacting directly with a ligand or ribozyme active site. These observations suggest that judicious placement of conformationally restricted nucleotides biased into the *syn* conformation could enhance RNA folding and catalysis. Such changes could also be useful for locking RNAs into functionally competent folds for use in X-ray crystallography and NMR.

Keywords: *syn* nucleobases; aptamer; riboswitch; ribozyme; ribosome

INTRODUCTION

RNA can occupy a vast number of conformational states, owing to the high number of rotatable bonds in the nucleotide (Duarte and Pyle 1998; Murthy et al. 1999; Bloomfield et al. 2000; Richardson et al. 2008). Indeed, flexibility of nucleotides is key to the complexity and diversity of RNA folding and to the many functions that RNA performs in the cell (Leontis et al. 2006; Parisien and Major 2008). Determining the most prevalent and functionally relevant conformations of the nucleotides and understanding their molecular basis form the foundations for predicting and manipulating RNA tertiary structure.

The number of RNA structures deposited in the Protein Data Bank has risen dramatically over the past decade, af-

fording the opportunity to survey and analyze tertiary structures for conformational biases that might be related to function. In the present study, we examine functional RNAs for occurrence, interactions, and functionality of nucleotides possessing the *syn* glycosidic conformation.

The most common orientation of the nucleotide in RNA is with the base rotated away from the sugar, in the *anti* conformation (Fig. 1). In the less common *syn* conformation, the base is rotated about the glycosidic bond to occupy the space closer to or above the furanose ring (Saenger 1984). The *syn* conformation affords a more compact form of the nucleotide, although it tends to be somewhat higher in energy and therefore more rare. Indeed, experiments and calculations indicate that the bases in most A-form RNA and B-form DNA helices are entirely in the *anti* conformation (Bloomfield et al. 2000), and hard-sphere potential calculations support rarity of the *syn* conformation (Murthy et al. 1999).

Importance of *syn* nucleobases in functional RNAs has been probed experimentally through incorporation of conformationally restricted nucleotides (CRNs). Bases such as 8-bromoguanosine (8-BrG) and 8-azidoadenosine (8-N₃A) (Dias et al. 1994; Maruta et al. 1998; Proctor et al. 2003;

¹Present address: Merck, Manufacturing Division, West Point, PA 19486-0004, USA.

²Present address: School of Pharmacy, University of Pittsburgh, Pittsburgh, PA 15213, USA.

³Corresponding author.

E-mail pcb5@psu.edu.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.2759911>.

FIGURE 1. Definition and statistics of *syn* nucleobases in RNA. (A) Classifications of nucleobases by glycosidic torsion angle χ . Bases are considered *syn* if $90^\circ \geq \chi \geq -90^\circ$, with the subcategorization of $90^\circ \geq \chi \geq -45^\circ$ being classified as full *syn* and $-45^\circ > \chi \geq -90^\circ$ being classified as intermediate *syn*. The other half of the unit circle consists of *anti* orientations. (B) *syn* nucleobase statistics found in functional RNA distributed among four sets of torsion angles. (C) Same as B but for the *anti* nucleobases in the control data set. The most populated region was that of $-135^\circ > \chi \geq -180^\circ$. (D) Examples of a purine in a full *syn* conformation, (E) an intermediate *syn* conformation, and (F) an *anti* conformation.

Yajima et al. 2007; Hammond et al. 2010), which force the base to take the *syn* orientation, have been shown to enhance function by one of two mechanisms: (1) favoring the native state or (2) disfavoring nonfunctional states (see Scheme 1).

- (i) In the *syn* state, favorable inter- or intramolecular interactions by the base occur, or;
- (ii) In the *anti* state, there is steric clash with another portion of the RNA, possibly in complex secondary and tertiary structures (Proctor et al. 2003).

Several examples can be given of where incorporation of a modified nucleotide altered a conformational equilibria toward a *syn*-containing state. In a G-quadruplex, incorporation of 8-BrG allowed a *syn*-containing quadruplex topology to be determined and thermostability to be enhanced (Dias et al. 1994). For the leadzyme, incorporation of 8-BrG at position 24 allowed for an increased rate of cleavage of the ribozyme by pre-organizing the binding pocket/active site (Yajima et al. 2007). Conversely, in small model systems, 8-BrG destabilized the duplex and promoted population of the hairpin state (Proctor et al. 2003) or populated a *syn* G conformer over a competing *anti* conformer in a duplex (Hammond et al. 2010). In addition, the bulky C8 substituent 8-hydroxyisopropyl drives adenosine

into *syn*, while butylamine at the C8 position favors the *anti* conformation through hydrogen-bonding of N8H to O5' (Saenger 1984).

On the basis of the above observations, it stands to reason that *syn* residues may be present in areas of other RNAs where the cost of adopting the conformation is offset by favorable interactions that are important to the overall fold of the RNA. *syn* nucleobases are therefore expected to be found at positions of functional importance within the RNA. These factors motivated the present bioinformatics study, wherein we analyze aptamer, riboswitch, ribozyme, and ribosomal structures for *syn* nucleobases using the program MC-Annotate (Gendron et al. 2001), as well as conduct downstream analysis of torsion angles and RNA function

SCHEME 1.

via the primary literature (for details, see Materials and Methods). Our results indicate that *syn* nucleobases are present in the vast majority of functional noncoding RNAs and that they participate in tertiary interactions and cluster in functionally important parts of the RNA.

RESULTS

Overview of *syn* nucleobase incidence

We surveyed four classes of functional RNAs for *syn* nucleobases: ribozymes, riboswitches, aptamers, and the ribosome (Table 1). Our hypothesis was that because the *syn* nucleobase is compact and generally oriented away from the helical axis, it may engage in tertiary interactions that enhance function. These particular RNAs were chosen because there are moderate- to high-resolution structures available for multiple representatives, as well as complementary experimental literature. In addition, with the exception of the ribosome, these RNAs function in the absence of proteins, suggesting structural and functional roles inherent to the RNA. Structures from each RNA class were analyzed for *syn* nucleobase occurrence, and any duplicate structures were

averaged to avoid overcounting. To be scored, crystal structures had to have a resolution of at least 3.1 Å, and all atoms in the *syn* residues had to have a temperature factor of 79 Å² or lower (equivalent to a mean square displacement of ≤ 1 Å). In addition, any *syn* nucleobases that were involved in crystal contacts were excluded. NMR structures had heavy atom RMSDs under 3 Å. In all, we surveyed 296 pdb files (Supplemental Table S1) and found that 51 fit the criteria for further analysis.

We found *syn* nucleobases in the majority of functional RNAs surveyed (Table 1). For example, 18/24 aptamers, 7/8 riboswitches, and 8/8 small ribozymes possess at least one *syn* nucleobase. In the case of large ribozymes, the group II intron has 18 *syn* nucleobases, and in the ribosome the average number of *syn* nucleobases in 23S, 16S, and 5S rRNAs is 136, 53, and 4, respectively.

In terms of total base count, *syn* nucleobases are relatively rare, comprising just 4% of all bases surveyed (Table 1). The fraction of *syn* nucleobases per nucleotide is largest for aptamers at 5.0% and smallest for small ribozymes at 3.4%. In the ribosome, *syn* frequency varies between the subunits and is 20% higher in the 50S subunit (4.2%) than the 30S subunit (3.5%).

Next, we focused on regions of the RNAs forming tertiary structure, defined as having tertiary stacking interactions and tertiary base-pairing. Interestingly, for these regions, the frequency of *syn* nucleobases is much higher: Of all tertiary interactions, nearly 1 in 8 contain at least one *syn* nucleobase, giving an overall tertiary structure density (TSD) of 12.3% (Table 1). We find that TSD is greatest in aptamers at 16.4% and smallest in riboswitches, still with a TSD of 9.0%. These observations reveal a strong enrichment of *syn* nucleobase frequency within tertiary structure of 2.4-fold to 3.5-fold (Table 1, column 5), which suggests that *syn* nucleobase occurrence may be driven by the need to form compact global folds.

We then divided *syn* nucleobases into two subcategories: “full *syn*” and “intermediate *syn*.” Each of these subcategories has unique structural and functional properties, as detailed in Tables 2 and 3. Intermediate *syn* is defined as having a χ torsion angle between -45° and -90° , and full *syn* as having an angle between $+90^\circ$ and -45° (Fig. 1A,D,E). Using these definitions, slightly more than half (56%) of all *syn* nucleobases are categorized as being of the intermediate *syn* type (Fig. 1B, fourth wedge).

TABLE 1. Whole molecule *syn* nucleobase statistics

RNA category	Number of unique structures surveyed	% <i>syn</i> nucleobases ^a	% of tertiary interactions ^b with <i>syn</i> nucleobases (TSD)	Enrichment ^c
All ^d	51	4.2% (305/7311)	12.3% (328/2660)	2.9
Aptamers	24	5.0% (38/762)	16.4% (38/232)	3.3
Riboswitches	8	3.7% (28/749)	9.0% (23/255)	2.4
Small ribozymes ^e	8	3.4% (15/437)	9.9% (13/131)	2.9
Group II intron	1	4.5% (18/397)	15.9% (24/151)	3.5
Ribosome ^f	2	4.2% (193/4550)	11.9% (212/1781)	2.8
23S	1	4.7% (136/2906)	14.4% (168/1168)	3.0
16S	1	3.5% (52/1504)	9.0% (60/665)	2.6
5S	1	3.3% (4/121)	7.5% (4/53)	2.3

^aNumber of *syn* nucleobases divided by total number of bases in structures for a given category. There were 289 *syn* nucleobases among the 51 representative structures, and 286 unique *syn* nucleobases. Of this unique subset, 238 were considered suitable for molecular level analysis carried forward into Tables 2 and 3. The full distribution in the *syn* database is 119 A (42%), 110 G (38%), 26 C (9%), and 31 U (11%). These can be subdivided into 129 full and 157 intermediate *syn* conformations. Among the full *syn* nucleobases, the distribution is 47 A (36%), 58 G (45%), 10 C (8%), and 14 U (11%); and for the intermediate *syn* nucleobases, the distribution is 72 A (46%), 52 G (33%), 16 C (10%), and 17 U (11%). Overall, there does not appear to be a bias toward full or intermediate *syn* for any particular nucleobase, as all full versus intermediate Z-scores were < 3 .

^bNumber of *syn* nucleobase-containing tertiary interactions divided by total number of tertiary interactions, which includes tertiary stacking and tertiary base pairs (as defined in Materials and Methods). This category is referred to in the text as Tertiary Structure Density, or TSD. These entries can be larger in value than the total number of *syn* nucleobases because some *syn* nucleobases have multiple tertiary interactions.

^cPercent tertiary interactions with *syn* nucleobases divided by percent *syn* nucleobases overall.

^dIncludes all RNAs below, as well as tRNA and miscellaneous other RNAs.

^eSmall ribozymes include hairpin, HDV, hammerhead, leadzyme, GlmS, and Diels-Alder.

^fThe % *syn* nucleobases is averaged over 25 structures. The tertiary interaction values were calculated from pdb 1K73 and 3OGE only. These pdb files were chosen on the basis of good resolution and occurrence of *syn* nucleobase frequency near the average of all ribosome structures.

TABLE 2. Sugar pucker orientations among *syn* nucleobases

Sugar pucker	Total <i>syn</i> ^c	Full <i>syn</i> ^{d,e}	Intermediate <i>syn</i> ^f	<i>anti</i>
C3'-endo-like ^a	32% (75/238)	48% (50/105)	19% (25/133)	86% (223/258)
C2'-endo-like ^b	68% (163/238)	52% (55/105)	81% (108/133)	14% (35/258)

^aC3'-endo-like includes C3'-endo, C4'-exo, C2'-exo, and C1'-endo sugar puckers (Saenger 1984).

^bC2'-endo-like is defined as C2'-endo, C4'-endo, C3'-exo, C1'-exo, and O4'-endo sugar puckers.

^cThe Z-score between total *syn* and *anti* was $Z = 12.3$. $Z > 3$ was considered significant.

^dThe Z-score between full *syn* and intermediate *syn* was $Z = 4.8$.

^eThe Z-score between full *syn* and *anti* was $Z = 10.8$.

^fThe Z-score between intermediate *syn* and *anti* was $Z = 13.0$.

We also composed a control set of *anti* nucleobases from the same structures (see Materials and Methods). Among this control data set, the majority of *anti* nucleobases (82%) have an angle between -135° and 180° (Fig. 1C, third wedge). Thus, the most populated *syn* and *anti* conformations differ by an $\sim 90^\circ$ rotation about the glycosidic bond.

Next, we analyze the distribution of *syn* nucleobases among the different nucleosides. Of the 286 unique *syn* nucleobase entries, 80% are purines (Table 1). For comparison, in the randomly chosen *anti* control, 56% of the bases were purines. Comparison of these *anti* and *syn* data sets give a Z-score of 6.3, indicating very significant preference for purines among *syn* nucleobases in functional RNAs. This observation also agrees with previous reports on nucleosides, where purines were found to be more common than pyrimidines among *syn* nucleobases (Saenger 1984). The bias against pyrimidines has been attributed to two factors: (1) a steric clash between the O2 and the ribose sugar, and (2) favorable hydrogen-bonding between the N3 of purines and the O5' of the ribose sugar (Saenger 1984).

The full distribution in the *syn* database is 119 A (42%), 110 G (38%), 26 C (9%), and 31 U (11%). These can be subdivided into full and intermediate *syn* conformations. Among the full *syn* nucleobases, the distribution is 47 A (36%), 58 G (45%), 10 C (8%), and 14 U (11%); and for the intermediate *syn* nucleobases, the distribution is 72 A (46%), 52 G (33%), 16 C (10%), and 17 U (11%). Overall, there does not appear to be a bias toward full or intermediate *syn* for any particular nucleobase ($Z < 3$).

The pucker of the ribose sugar is typically linked to the conformation of the nucleobase (Saenger 1984). We determined the sugar pucker of each base and sorted these into two bins, "C3'-endo-like" and "C2'-endo-like," using the rules provided in Table 2. The distribution of sugar puckers for the *syn* nucleobases was found to be significantly different from for the *anti* control ($Z = 12.3$). For the *anti* database, the sugar pucker is almost exclusively C3'-endo-like (86%), as expected. For the *syn* database, we found an overall partial bias for the other pucker, C2'-endo-like, at 68%. However, subdividing the *syn* nucleobases into full and intermediate *syn* revealed that intermediate *syn* is more

strongly associated with C2'-endo-like pucker: For intermediate *syn*, 81% of all sugars are C2'-endo-like, while for full *syn* there are approximately equal numbers of nucleobases in each subcategory. The reason for this difference likely has to do with steric repulsion. When the nucleobase assumes the intermediate *syn* conformation, a steric clash occurs between the H2' of a C3'-endo ribose sugar and either the H6 of a pyrimidine or the H8 of a purine, which is relieved by the C2'-endo like pucker (Fig. 2).

Notably, this significant structural difference justifies our division of *syn* nucleobases into intermediate and full *syn* subcategories. It is interesting to note that substitution of C6 with N in 6-azauridine, which removes the H6 proton, allows population of the intermediate *syn* conformation (Schwalbe and Saenger 1973; Saenger 1984).

We then investigated whether neighboring sequence influences adoption of the *syn* conformation. Nearest neighbors of each *syn* nucleobase were recorded and information content was calculated, as described in the Materials and Methods, with 0 bits being defined as no certainty and 2 bits being absolute certainty. As shown in Supplemental Figure S1, information content was near zero (<0.25) for all *syn* nucleobase-nearest-neighbor combinations. Overall, it appears that primary structure plays little or no role in determining *syn* nucleobase position or identity. This finding is in agreement with the notion that *syn* nucleobases participate in tertiary interactions rather than local interactions, as described below.

TABLE 3. Stacking and base-pairing interactions of *syn* versus *anti* nucleobases

	<i>syn</i> nucleobases	<i>anti</i> nucleobases
Stacking interactions ^a	77% (184/238)	87% (225/258)
Tertiary stacking ^b	74% (136/184)	34% (77/225)
Base-pairing ^c	57% (136/238)	81% (209/258)
Tertiary base-pairing ^d	93% (127/136)	36% (75/209)

^aComposed of secondary and tertiary stacks (as defined in Materials and Methods) among the 238 unique *syn* nucleobases of sufficient quality. The Z-score between *syn* and *anti* was $Z = 2.9$. Of the 184 stacks, 68 do so without base-pairing. $Z > 3$ was considered significant.

^bThe percent *syn* nucleobase stacks defined as tertiary. The Z-score between *syn* and *anti* was $Z = 9.0$.

^cComposed of secondary and tertiary base-pairing (as defined in Materials and Methods) among the 238 unique *syn* nucleobases. The Z-score between *syn* and *anti* was $Z = 5.8$.

^dThe percent *syn* nucleobase pairs defined as tertiary. The Z-score between *syn* and *anti* was $Z = 10.5$. The value for *anti* nucleobases is not smaller because *anti* nucleobases were chosen randomly from the same set of functional RNAs as the *syn* nucleobases and so had the possibility for tertiary interactions.

FIGURE 2. Sterics as a driving force for intermediate *syn* nucleobases to adopt the C2'-endo-like conformation. Shown in space-filling is cytidine monophosphate in the intermediate *syn* conformation for (A) C2'-endo and (B) C3'-endo sugar pucker, and adenosine monophosphate in the intermediate *syn* conformation for (C) C2'-endo and (D) C3'-endo sugar pucker. Note that the H6 and H2' of cytidine, and the H8 and H2' of adenosine clash when the sugar is C3'-endo (depicted with double-headed arrows in *right-hand panels*) but not when it is C2'-endo (*left-hand panels*).

Involvement of *syn* nucleobases in tertiary structure

On the basis of the above observations that *syn* nucleobases are enriched in tertiary structure, we analyzed frequency of the *syn* conformation in tertiary structure stacking and base-pairing (for details, see Materials and Methods). First, we investigated stacking of the *syn* nucleobases (Table 3). Most *syn* nucleobases (77%) engage in some type of stacking, either secondary or tertiary, with 74% of all *syn* nucleobase stacks being nonadjacent, or tertiary stacks. Lower occurrence of adjacent stacks (39% or 79/184) can be attributed to unsuitable orientation of the nucleobases between neighbors when at least one of the nucleobases is *syn*.

Next, stacking of the *anti* nucleobases was compared. Similar to the *syn* nucleobases, most *anti* nucleobases (87%) engage in some type of stacking. However, just 34% of *anti* stacks are tertiary stacks, which is less than half the percentage in the *syn* database. Instead, *anti* nucleobases interact largely through adjacent stackings (91% or 204/225), as expected because many of these nucleobases come from helical regions of RNA.

Next we analyzed base-pairing involving *syn* nucleobases (Table 3). Across the database, 57% of all *syn* nucleobases engage in some type of base-pairing, either secondary or tertiary; however, a remarkable 93% of all *syn* nucleobase pairs are tertiary base pairs (defined in Materials and Methods). This is again consistent with *syn* nucleobases being crucial components of RNA tertiary architecture. There were no significant trends with base-pair type with the purines. For example, GC, AU, and GU pairs of all types were found, and these had similar frequencies as AG, GG, and AA pairs. AC pairs, however, were rarely observed with *syn* As.

A comparison to *anti* nucleobases was made for base-pairing as well. Here, even more nucleobases engaged in some type of base-pairing (81%), as expected because many

of these nucleobases come from helical regions. However, only 36% of these base pairs were of the tertiary type, which is just a third the number in the *syn* database. In conclusion, the vast majority of *syn* nucleobase interactions involve tertiary stacking and tertiary base-pairing.

Analysis of *syn* nucleobases in different classes of functional RNA

The preceding sections illustrated that *syn* nucleobases are enriched within tertiary interactions and that *syn* nucleobase stacks and base pairs are overwhelmingly tertiary in nature. These observations suggested that *syn* nucleobases might play important roles in the active sites of functional RNAs. To test this hypothesis, we examined positioning of the *syn* nucleotides within individual aptamers, riboswitches, ribozymes, and ribosomal RNA structures.

Aptamers and riboswitches

Aptamers and riboswitches are RNAs that bind small molecules or proteins with high specificity and sensitivity. Aptamers are RNAs isolated from random libraries by *in vitro* selection techniques (Osborne and Ellington 1997), while riboswitches are naturally occurring aptamers that also contain a signal-transduction, or expression, domain (Tucker and Breaker 2005). We begin with analysis of *syn* nucleobase statistics (Table 4) and then move on to various aptamer and riboswitch structures (Fig. 3).

We found that *syn* nucleobases are plentiful in both aptamers and riboswitches. Almost all riboswitch structures in the PDB (7/8) have at least one *syn* nucleobase, while 75% (18/24) of unique aptamer structures have at least one *syn* nucleobase. We then analyzed the fraction of aptamer- and riboswitch-derived *syn* nucleobases that play a functional role in the RNA as discerned by analysis of the original literature on each aptamer, and determined those that are found either within the binding pocket, or those that contact the ligand directly (Table 4). The majority (80%) of *syn* nucleobases in RNA aptamers play some functional role. In addition, among aptamers 91% of functional *syn* nucleobases are found in the binding pocket, with 63% directly hydrogen-bonding or stacking with the ligand. In riboswitches, 63% of all *syn* nucleobases contribute to function, but in contrast to aptamer data, only 44% are in the binding pocket, and just 25% interact directly with the target ligand. Thus, in riboswitches, *syn* nucleobases are used in functions beyond ligand binding: The remainder of the functional *syn* nucleobases engage in tertiary interactions separated in space from the aptameric domain and appear to be important to the conformational switching and gene regulatory portions of the riboswitch (see below). This conclusion is reinforced by the data in Table 1, which indicate that 5% of all nucleobases are *syn* in aptamers, as compared to just 3.7% in riboswitches, and that 16.4% of all tertiary

TABLE 4. Functional roles for *syn* nucleobases in RNA aptamers compared to riboswitches

	Total <i>syn</i> nucleobases	Full <i>syn</i> nucleobases	Intermediate <i>syn</i> nucleobases
Aptamers^a			
Functional role (direct or indirect)	80% (35/44) ^d	71% (15/21)	87% (20/23)
<i>syn</i> nucleobase in binding pocket	91% (32/35) ^e	100% (15/15)	85% (17/20)
<i>syn</i> nucleobase with direct ligand contact ^b	63% (22/35) ^f	73% (11/15)	55% (11/20)
Riboswitches^c			
Functional role (direct or indirect)	63% (32/51)	52% (13/25)	73% (19/26)
<i>syn</i> nucleobase in binding pocket	44% (14/32)	31% (4/13)	53% (10/19)
<i>syn</i> nucleobase with direct ligand contact	25% (8/32)	15% (2/13)	32% (6/19)

^a18/24 aptamers have at least one *syn* nucleobase.^bThese are either stacking or hydrogen-bonding interactions.^c7/8 riboswitches have at least one *syn* nucleobase.^dThe Z-score between aptamers and riboswitches was $Z = 1.5$. $Z > 3$ was considered significant.^eThe Z-score between aptamers and riboswitches was $Z = 4.5$.^fThe Z-score between aptamers and riboswitches was $Z = 3.6$.

interactions contain *syn* nucleobases in aptamers, as compared to just 9.0% in riboswitches. Lastly, a comparison between full and intermediate *syn* nucleobases is provided in Table 4 as well and shows that these two types of *syn* nucleobases provided similar functionality to the RNAs.

Next, we consider specific aptamer and riboswitch structures with regard to function. Of all the *syn* nucleobases identified in aptamers and riboswitches, ~70% are classified in the original literature as being relevant to the molecule's inherent function. Various themes regarding these functional roles can be discerned from inspection of structures. In aptamers, *syn* nucleobases cluster in binding pockets and make direct contacts to the ligands (Fig. 3A–C). For the citrulline aptamer (Fig. 3A), *syn* nucleobases G29, G30, and G35 comprise three of the eight nucleobases that directly contact the ligand by hydrogen-bonding (Yang et al. 1996). In addition, C28, G29, and G30 form a three-base network of *syn* nucleotides at the binding site. In the malachite green aptamer (Fig. 3B), *syn* nucleobases G24, G29, and A31 form essential portions of the binding pocket, with G24 and G29 directly stacking on the malachite green ligand (Baugh et al. 2000). In the ATP aptamer (Fig. 3C), *syn* nucleobases again play a prominent role, with one-third of the binding pocket being *syn* (A9, A12, and G30) (Dieckmann et al. 1996). We also

note that *syn* nucleobases occur in non-functional regions of this structure too, including U23 and G25 in the UUCG tetraloop as expected from NMR studies on this tetraloop (Cheong et al. 1990). While most aptamers have *syn* nucleobases in their binding motifs, some do not, with the most notable examples being the theophylline and caffeine aptamers (Zimmermann et al. 2000), although the possibility of transient *syn* conformations remains possible (see Discussion).

As mentioned, riboswitches contain both aptameric and expression domains, so here non-ligand-binding roles for *syn* nucleobases become possible (Tucker and Breaker 2005). Interestingly, several riboswitches have *syn* nucleobases present both in the binding pocket and in long-range tertiary interactions, which are important for the expression domain (Fig. 3D–F).

In the guanine riboswitch (Fig. 3D), the binding pocket is formed with *syn* A23, while a loop/loop interaction is mediated by *syn* A65, which is important in the global fold of the riboswitch according to Batey et al. (2004). In the c-di-GMP riboswitch (Fig. 3E), full *syn* A47, one of two *syn*

FIGURE 3. Location of *syn* nucleobases in aptamers and riboswitches. The *syn* nucleobases (blue); ligands (red). (A) The citrulline aptamer (1KOD) with *syn* nucleobases C28 (-60°), G29 (-67.8°), G30 (31.2°), and G35 (2.3°). The χ angles are provided in parentheses. (B) The malachite green aptamer (1Q8N) with *syn* nucleobases G24 (-64.1°), G29 (-60.1°), and A31 (-54.6°). (C) The ATP aptamer (1RAW) with *syn* nucleobases A9 (-82.3°), A12 (-63.3°), U23 (39.3°), G25 (53.7°), and G30 (81.2°). (D) The guanine riboswitch (1U8D) with *syn* nucleobases A23 (83.2°) and A65 (-71.4°). (E) Cyclic-di-GMP riboswitch (3MXH) with *syn* nucleobases A43 (66.9°) and A47 (-84.7°). (F) The lysine riboswitch (3D0U) with *syn* nucleobases G8 (38.8°), G9 (-63.8°), C10 (41.2°), C38 (69°), G77 (-73.2°), A123 (-11.4°), and U125 (-41.2°).

nucleobases, intercalates between the two guanine rings of the ligand and is the most conserved nucleotide in the riboswitch (Smith et al. 2010). The lysine riboswitch (Fig. 3F) contains seven *syn* nucleobases, with four (G8, G9, C10, and G77) clustered in the binding pocket (Garst et al. 2008) and two more (A123 and U125) present in the region where the P4 pentaloop docks to P2/P2a (bottom of the structure). Within the binding pocket, the rings of G8 and G77 contact methylene groups of the ligand, playing key roles in binding specificity, while the exocyclic amine of G9 and the 2'-OH of G8 hydrogen-bond to the carboxylate of lysine. The O4' of G77 and its nonbridging phosphate oxygen also hydrogen-bond to the ϵ NH₂ group. In sum, *syn* nucleobases play critical roles in ligand binding and tertiary interactions in this class of RNAs.

Ribozymes

Ribozymes are catalytic RNAs that are both found in nature and isolated from selection experiments (Fedor 2009). It is customary to divide ribozymes into small ribozymes, which leave 2',3'-cyclic phosphate and 5'-hydroxyl termini, and large ribozymes, which leave 2',3'-hydroxyls and 5'-phosphate termini. As we show here, both small and large ribozymes feature *syn* nucleobases in their active sites. We begin with analysis of several small ribozymes, then move to the larger ribozymes.

The glucosamine-6-phosphate (glmS) ribozyme is a riboswitch-ribozyme (Fig. 4A), which occurs in the 5' UTR of the gene that codes for the glucosamine synthetase enzyme (Barrick et al. 2004; McCown et al. 2011). High concentrations of glucosamine-6-phosphate (GlcN6P) bind the 5' UTR and induce the RNA to self-cleave, which silences gene expression. The most notable *syn* nucleobase in this ribozyme is G1 (Klein and Ferre-D'Amare 2006; Cochrane et al. 2007). This nucleotide, which supplies the leaving group in the reaction, hydrogen-bonds to GlcN6P, helping to position it in the active site. The 2-amine of GlcN6P appears to act as a general acid to protonate the 5'-bridging oxygen of G1 (Klein and Ferre-D'Amare 2006). In addition, G1's position is stabilized by stacking with A35, which itself is *syn*.

Next, we examined the lead-dependent ribozyme (or "leadzyme") (Fig. 4B). Differing structures of the leadzyme have been reported by NMR, crystallography, and computational studies, each of which contained a unique *syn* nucleobase (Hoogstraten et al. 1998; Lemieux et al. 1998; David et al. 2001; Wedekind and McKay 2003). Yajima et al. (2007) made a series of synthetic RNA molecules with the *syn* nucleobase 8-bromoguanosine (8BrG) base incorporated at each site of interest. (The bulk of the bromine precludes the normal *anti* conformation for 8BrG.) The highest rate of cleavage—up to 30-fold faster than wild-type—was observed for 8BrG24, consistent with the computational structure (Lemieux et al. 1998), which had *syn* G24 (Fig.

FIGURE 4. Location of *syn* nucleobases in ribozymes. (Red arrows) Cleavage sites. (A) The glmS ribozyme (PDB 2HO7) has four *syn* nucleobases, including G1 (-65.4°) and A35 (-88.0°) in the active site. The cofactor is shown in red. (B) The leadzyme (PDB coordinates at <http://www.lbit.iro.umontreal.ca/structures/leadzyme.pdb>) has three *syn* nucleobases, including G24 (-23.7°), A25 (-18.6°), and G26 (59.8°) in the active site. The two red spheres represent lead ions. (C) The hairpin ribozyme (PDB 1M5K) has three *syn* or nearly *syn* nucleobases, including G1 (93.7°) and A38 (71.3°) in the active site. (D) The HDV ribozyme (PDB 3NKB) has three *syn* nucleobases, including G25 (71.9°) in the active site and G76 (-85.3°) adjacent to the catalytically critical C75. (E) The hammerhead ribozyme (PDB 2GOZ) has three *syn* nucleobases G8 (-85.6°), AL3 (-88.6°), and AL6 (-70.4°). (F) The group II intron (PDB 3IGI) with 18 *syn* nucleobases, which appear to cluster at key tertiary junctions such as the coordination loop and the five-way junction. See text for full description of functional roles of the *syn* nucleobases in these structures.

4B). In this conformation, G24 is in close proximity to two Pb²⁺ ions, positioning them for bond scission. In addition, two other active-site residues, A25 and G26, are intermediate *syn* and help form the Pb²⁺ binding pocket, with G26 engaged in a reverse Watson-Crick base pair (Lemieux et al. 1998).

The hairpin ribozyme (Fig. 4C) also has *syn* nucleobases at key positions in the active site (Rupert et al. 2002). Residue A38 is *syn* and has been implicated as the general acid in the cleavage reaction (Bevilacqua 2003; Wilson and Lilley 2011). Structures with mutants of A38 showed that when a variant residue at position 38 takes on the *anti*

conformation, local structure is disrupted and self-cleavage is inhibited (Macelrevey et al. 2008). In addition, the leaving group residue G1 is *syn*, and it stacks on the *syn* A38 general acid. Thus, both the hairpin and glmS ribozymes have a *syn* nucleobase at G1 that stacks on a *syn* A. Supplemental Figure S2 shows the structural overlay of the two motifs in which it can be seen that the six-membered rings of the two purines stack to provide stability.

The hepatitis delta virus (HDV) ribozyme uses a *syn* nucleobase in its active site (Fig. 4D). Residue G25 adopts the full *syn* conformation, where it forms a rare reverse G•U wobble with U20 (Chen et al. 2010). Like a normal G•U wobble, this reverse wobble presents a negative potential for binding metal ions, including a partially hydrated Mg^{2+} , which is enhanced by positioning of phosphates (Veeraraghavan et al. 2011). This highly negatively charged motif has also been implicated in providing the driving force for shifting of the pK_a of C75 toward neutrality. Mutation of the reverse G•U wobble to an AU resulted in a more than a 1000-fold loss in reaction rate (Tanner et al. 1994), indicating that this highly unusual motif is crucial to ribozyme function. Moreover, the wobble has been found in all kingdoms of life (Webb et al. 2009), supporting its importance. In addition, G76, which follows the catalytically critical C75, is *syn* and extruded, which promotes stacking of C75 on A77.

Bases with the *syn* nucleotide conformation are also found in the full-length native hammerhead ribozyme, where they are involved both in the active site and in mediating loop–loop tertiary interactions (Fig. 4E; Martick and Scott 2006), much like in the riboswitches. In the active site, G8 is *syn*, where its 2'-OH group serves as the putative general acid. In addition, *syn* nucleobases AL3 and AL6 mediate long-range loop–loop interactions.

Next, we turn to large ribozymes. Remarkably, the group I intron, in three structures with ribose at the ω G position, does not show any functionally significant *syn* nucleobases (Guo et al. 2004; Golden et al. 2005; Stahley and Strobel 2005). An intermediate *syn* nucleobase is present in the group I structure prior to the second step of splicing, but this structure had an inactivating deoxyribose at ω G (Adams et al. 2004).

In contrast, the group II intron structure (Fig. 4F) has numerous *syn* nucleobases (18 of 397 nucleotides, or 4.5%) within its structure, and a tertiary structure density of 15.9% (Table 1; Toor et al. 2008). Although none of the *syn* nucleobases occur at splice sites, the *syn* nucleobases cluster, often at key tertiary structure junctions. In particular, *syn* A35 partakes in the T-loop stack portion of the five-way junction, the keystone of the intron's tertiary structure. In addition, *syn* nucleobases A137, A138, and A51 engage in key interactions at the interface between the coordination loop and the κ -extension region. In sum, *syn* nucleobases are important both to RNA catalysis and folding in catalytic RNAs.

50S and 30S ribosomes

The ribosome is an RNP in which RNA plays extensive structural and chemical roles (Steitz and Moore 2003; Korostelev et al. 2008). Overall, ribosomal RNA has statistics similar to the total *syn* database with regard to *syn* nucleobase properties, such as frequency, tertiary interaction, and enrichment (Table 1). Breaking rRNA out into the different rRNAs, 23S rRNA has the highest *syn* nucleobase frequency among all naturally occurring functional non-coding RNAs, at 4.7%, while 16S and 5S rRNAs have 3.5% and 3.3% *syn* nucleobases, respectively. Moreover, the TSD is approximately twofold higher in 23S rRNA (14.1%) than in 16S and 5S rRNAs, at 9.0% and 7.5%, respectively. These observations have implications for the catalytic function of the 23S rRNA (see Discussion).

Figure 5 reveals that *syn* nucleobases cluster within both 23S and 16S rRNA, including Domains IV, V, and the L7/L12 stalk of Domain II of 23S rRNA. Domain IV, which forms the 50S subunit's half of the interaction face with the 30S, has 14 bases *syn*, and they are oriented along this binding face, suggestive of an importance in quaternary structure. There is also a high density of *syn* nucleobases in Domain V of 23S rRNA, the key region for peptidyl transferase activity, with 18 *syn* nucleobases. In addition, the *syn* nucleobases in these clusters appear frequently at strongly conserved residues (Supplemental Table 4), indicating that these positions are relevant to ribosomal function. It thus appears that *syn* nucleobases in the ribosome make important

FIGURE 5. Location of *syn* nucleobases in the 50S (1K73) and 30S (3OGE) subunits of the ribosome. (A) The 50S subunit showing the positions of the *syn* nucleotides. Domain IV *syn* nucleobases (red); Domain V *syn* nucleobases (magenta); L7/L12 stalk *syn* nucleobases (green); all other *syn* nucleobases (blue). RNA is in wire, and protein is not shown. (B) Depiction of the 50S subunit from A with the wire removed, showing just the *syn* positions of the 23S rRNA and 5S rRNA. (C) The 30S subunit showing the positions of *syn* nucleotides (blue). RNA is in wire, and protein is not shown. (D) Depiction of the 30S subunit from C with the wire removed, showing just the *syn* positions of the 16S rRNA.

contributions to both folding and catalysis, just like in the ribozyme, riboswitch, and aptamer structures.

Uncomplexed tRNA shows only a single *syn* nucleobase. Residue A9, which is in the joining region between the acceptor and D stems, is in an intermediate *syn* orientation in PDB 1EHZ, with a χ angle of -70.5° (Shi and Moore 2000). This residue is buried in the core of the tRNA, where it makes tertiary base-pairing interactions with another A, as well as tertiary stacking interactions, again illustrating the importance of *syn* residues to RNA folding and tertiary structure. Survey of 10 other tRNA structures in the PDB with an A at position 9 revealed that in eight cases, A9 is intermediate *syn*, and that in the other two, it is nearly so. Thus, *syn* at this position appears to be important for the tertiary fold of tRNA.

DISCUSSION

Recent efforts aimed at analyzing RNA structural information have focused on identifying and characterizing key structural motifs (Duarte and Pyle 1998; Gendron et al. 2001; Leontis et al. 2006; Abraham et al. 2008; Richardson et al. 2008; Xin et al. 2008). These studies have examined overall backbone conformation as well as hydrogen-bonding and stacking patterns among various RNA structures, but have not focused on *syn* versus *anti* conformations in those molecules. As presented herein, the majority of functional RNAs with tertiary structures have *syn* nucleobases in key regions. This occurs despite the *syn* nucleobase typically being higher in free energy than the *anti* conformation: For free nucleotides, the *syn* conformation for purines is slightly disfavored by ~ 1.2 kcal/mol in ΔG at 35°C , and more so for pyrimidines (Hemmes et al. 1974; Saenger 1984; Bloomfield et al. 2000). In helical regions, *syn* nucleobases are very strongly disfavored, by ~ 2.4 kcal/mol, owing to helix disruption (Proctor et al. 2003). However, the *syn* conformation can be populated, either by stabilizing it through hydrogen-bonding and stacking, or by steric repulsion of the *anti* form (Scheme 1). These mechanisms require tight packing and unusual geometries such as those present in RNA tertiary structures.

In functional RNAs, $\sim 3\%$ – 5% of the bases are in the *syn* conformation (Table 1), with approximately half of these being intermediate *syn* and half being full *syn* (Fig. 1B). The intermediate *syn* nucleobases prefer the C2'-endo-like sugar pucker, as opposed to the standard C3'-endo-like pucker present in *anti* nucleobases, while the full *syn* nucleobases did not show a sugar pucker preference (Table 2). Sugar pucker preferences could be understood in terms of steric clash between H2' of the sugar and H6 (pyrimidine) or H8 (purine) of the base, as illustrated in Figure 2.

The *syn* nucleobase conformation, when it occurs, is typically present in tertiary structures. This is best represented by the tertiary structure enrichment factor in Table 1, which ranges between 2.4 and 3.5. This observation,

together with the structural characteristic that *syn* nucleobases reach away from the inside of helices to interact with other parts of the RNA and help compact the nucleotide, suggested that *syn* nucleobases could play important functional roles. Inspection of the local environment of *syn* nucleobases in functional RNAs revealed that most *syn* interactions are either tertiary stacking (74% of all *syn* nucleobase stacks are tertiary) or tertiary base-pairing (93% of all *syn* nucleobase pairs are tertiary). Given these statistics, we examined individual RNA molecules for roles of *syn* nucleobases.

Inspection of aptamer and riboswitch structures revealed multiple *syn* nucleobases in the ligand-binding domain (Fig. 3), wherein *syn* nucleobases interact extensively with the ligand (Table 4). The citrulline, malachite green, and ATP aptamers use nearly all of their *syn* nucleobases in engaging the ligand; moreover, a majority of the binding site interactions are contributed by *syn* nucleobases (Fig. 3). The *syn* nucleobases help kink the backbone and position the bases in ways that create clefts to bind ligands. This theme is also present with clefts that bind metal ions in the leadzyme and HDV ribozymes: In the leadzyme, three *syn* nucleobases cluster where they engage Pb^{2+} ions in binding, while in the HDV ribozyme G25 is *syn*, which facilitates binding of various mono and divalent ion species (Veeraraghavan et al. 2011).

Relative to aptamers, riboswitches were found to have a smaller percentage of bases as *syn* in terms of TSD (Table 1). This appears to occur, in part, because riboswitches have greater size than aptamers. Moreover, in addition to binding a ligand, riboswitches transduce binding into changes in gene expression. The higher TSD in RNA aptamers likely reflects their simplicity: *syn* nucleobases may be a motif by which simple domains gain a high level of complexity without involving interactions with another strand. Riboswitches nonetheless still display a significant occurrence of *syn* nucleotides both within the binding pocket and at locations involved in tertiary structure packing (Fig. 3D–F).

The active sites of catalytic RNAs, like those of aptamers, contain multiple *syn* nucleobases. In the glmS and hairpin ribozymes, the leaving group nucleotide G1 is *syn* and is stacked upon by a *syn* A, which in the case of the hairpin serves as the general acid (A38) (Fig. 4A,C). It appears that the *syn* nature of the base is critical for orienting in-line attack by the 2'-hydroxyl in these cases, as well as for leaving group orientation. In fact, mutants of A38 in the hairpin ribozyme that lead to loss of the *syn* conformation induce inactive conformations by removing the imino proton from the 5'-bridging oxygen (Macelrevey et al. 2008). Thus, in the same way that *syn* nucleobases kink the backbone and create binding clefts for ligands and ions, they also create clefts that orient atoms for RNA catalysis.

The final theme that we observed with *syn* nucleobases is clustering. Clustering is observed in the binding sites of aptamers and ribozymes and is present on a larger scale in

the group II intron and ribosome. In the group II intron, the interface between the coordination loop and κ -extension reveals multiple *syn* nucleobases, as do other regions of the intron.

The large subunit rRNA, specifically Domains IV, V, and portions of II, evolved very early, significantly before the small subunit, and could be considered a molecular fossil dating back to the “RNA world” (Fox 2010). Most strikingly, the peptidyl transferase center, present in Domain V, is composed entirely of RNA, with no proteins within 17 Å of the active site, and remains catalytically active after extraction of most proteins (Noller et al. 1992); conversely, the decoding site of the 30S is intrinsically dependent on the constituent proteins (Smith et al. 2008). The 50S subunit also has a smaller proportional protein content than the 30S. The evolutionary age, lesser dependence on proteins, and catalytic function of 50S, arguably put more structural and chemical burden on the 23S rRNA and thus require a greater degree of structural complexity and compaction. This may be the basis for its increased utilization of *syn* nucleotides.

It is curious to note that the group I intron does not show any functionally significant *syn* nucleobases. These RNAs are generally very good folders: They fold rapidly and readily crystallize in functionally relevant structures (Sclavi et al. 1998; Guo et al. 2004; Golden et al. 2005; Stahley and Strobel 2005). These observations raise the possibility that *syn* nucleobases may render an RNA slow to fold or more prone to misfolds. Such observations have previously been made on pseudoknots, wherein intermediates can engage in alternative interactions that cannot form native tertiary structure (Chadavada et al. 2002).

There are other manners in which *syn* bases might present themselves. Recently, Al-Hashimi and colleagues demonstrated that base pairs in a free canonical DNA double helix undergo transient flipping to a Hoogsteen geometry in which the purine adopts the *syn* conformation (Nikolova et al. 2011). These observations raise the possibility that such fluctuations might be possible in RNA as well, which may further enhance the importance of the *syn* conformation, or might lead to *syn* nucleobases in RNAs with relatively few or no *syn* nucleobases in crystal structures, such as the group I intron or the theophylline and caffeine aptamers. In addition, *syn* nucleobase conformation can be conferred by damage to the DNA (and presumably RNA) such as alkylation or oxidation in 1-methyladenine or 8-oxoguanine (Beard and Wilson 2005; Brieba et al. 2005; Yang et al. 2008), or binding of proteins (Kitayner et al. 2010).

In conclusion, *syn* nucleobases are common in the active sites and tertiary folds of functional RNAs. Regions of *syn* nucleobases correlate with functional domains, and it remains possible that occurrence of *syn* geometries could be further populated by fluctuations of the base, or damage or modification of the RNA, as well as interaction with

proteins. Experimental methods to identify bases as *syn* could help in determining functional regions of RNAs; moreover, constraining of bases to the *syn* conformation, such as through judicious incorporation of bulky functional groups at the 8-position, could help enhance RNA folding and catalysis as already demonstrated for a quadruplex, model helices, and a small ribozyme (Dias et al. 1994; Proctor et al. 2004; Yajima et al. 2007; Hammond et al. 2010). Such changes might also lock RNAs in functionally competent folds for use in X-ray crystallography and NMR.

MATERIALS AND METHODS

Definition of *syn*

The *syn* conformation is defined by the IUPAC designation of a glycosidic torsion angle of $0^\circ \pm 90^\circ$ (Saenger 1984). For this study, we subdivided the *syn* nucleobases into full *syn* ($+90^\circ \geq \chi \geq -45^\circ$), where the base is directly above the sugar, or intermediate *syn* ($-45^\circ > \chi \geq -90^\circ$), where it is partially over the sugar (Fig. 1). Intermediate *syn* has in the past also been termed “high *anti*” (Saenger 1984).

Assembly of the *syn* database

Structures for analysis were obtained via the RCSB Protein Data Bank (Bernstein et al. 1977) by searching with the following terms: “RNA aptamer,” “ribozyme,” “riboswitch,” “tRNA,” “intron,” and “ribosome.” Supplemental Table 1 lists every PDB file analyzed in assembling the *syn* database, along with its resolution and class. To ensure adequate resolution to resolve glycosidic bond conformation and bonding interactions, only structures with a resolution of 3.1 Å or better were chosen for final analysis; 3.1 Å was chosen to allow discussion of the group II intron structure (Toor et al. 2008). Because of these criteria, RNase P in complex with its tRNA substrate was excluded from the analysis (Reiter et al. 2010). Our search is exhaustive through February 2011.

Candidate structures were downloaded as *pdb.gz* files and analyzed with the program MC-Annotate to find candidate *syn* nucleobases; MC-Annotate is a web-based program for analyzing RNA conformations based on the more extensive MC-Sym software package (Gendron et al. 2001). Exact torsion angles were measured using the program DSViewerPro (Accelrys), which in addition to providing detailed structural information served in every case to confirm the MC-Annotate output. Values of these angles are provided in Supplemental Table 2. All candidate bases were also screened for their temperature factors with Pymol (Schrodinger), and a temperature factor of $\leq 79 \text{ Å}^2$ (a mean square displacement of $\leq 1 \text{ Å}$) was required for the candidate base to be scored. NMR structures had heavy atom RMSDs under 3 Å.

MC-Annotate also provided definitions of tertiary interactions, glycosidic conformation, sugar pucker, stacking, and base-pairing data, which were used in the analysis. The program defined secondary and tertiary structure stacking as between adjacent bases and nonadjacent bases, respectively. Tertiary base-pairing and stacking were defined as arising from the interactions of separate helical regions (Gautheret et al. 1993; Major 2003; Parisien and Major 2008). Due to the often difficult nature of resolving sugar puckers (Keating and Pyle 2010), the results of MC-Annotate for sugar

puckers were verified by the use of the MolProbity program (Davis et al. 2004, 2007). If the MolProbity results disagreed with the MC-Annotate classification of sugar pucker, then the *syn* nucleotide entry was discarded, which was necessary in a only a few instances. Extruded bases that were not interacting intramolecularly, some of which were involved in crystal contacts, were excluded from the database.

Direct ligand contact was scored when the *syn* nucleobase either hydrogen-bonded to or was stacked upon a ligand in aptamers and riboswitches. Hydrogen-bonding was judged by use of the H-Bond Monitor tool in DSViewerPro, while stacking was assigned on the basis of a distance of ≤ 4 Å between the base and an aromatic moiety on the ligand. Functional roles for *syn* nucleobases were assigned by consulting prior biochemical studies in the primary literature.

In most instances, the PDB contains multiple structures for a single functional RNA. For instance, the purine riboswitch has 16 complete structures, while the hairpin ribozyme has 40 complete and partial structures deposited in the PDB. It was therefore necessary to implement a method to avoid overcounting: For whole molecule *syn* nucleobase frequency, a representative structure for each functional RNA was chosen generally using the structure that appeared most relevant in terms of in vivo function (e.g., cognate ligand bound, closest to wild-type sequence, and biological cations present). If multiple structures fit these criteria, then statistics were calculated from all structures, averaged, and weighted as one entry. For example, the streptomycin-bound RNA aptamer has two structures available in the PDB: 1NTA, which has two *syn* nucleobases, G12 and C18; and 1NTB, which has three *syn* nucleobases, G12, C18, and A8. Thus, out of five raw *syn* nucleobase database entries, three were considered for analysis: G12 and C18, which are identical with respect to the two structures, and A8 from 1NTB. In the case of two entries having the same sequence context and sugar pucker but different stacking interactions and base-pair type, statistics for sugar pucker and nearest neighbor contained one entry, while stacking and base-pairing statistics had two entries. We also compiled an *anti* database as a control, which came from functional RNAs and included both helical and nonhelical regions (Supplemental Table S3).

Most structures analyzed were crystal structures; however, we also considered 17 NMR structures. The majority of the NMR structures analyzed were of aptamers because, given their smaller size, more NMR solution studies have been conducted on aptamers than other complete RNAs. The NMR structure PDB files, which typically contain 10 or more separate structures, were divided into separate files for each structure using in-house code and analyzed individually. The average of all the χ angles of that residue in all files was used as the one χ angle data for a *syn* nucleobase. Sugar pucker data for *syn* nucleobases from NMR structures were excluded from analysis due to variability among the structures in a PDB entry. Tertiary interactions that remained consistent across the ensemble of conformers were retained as a single entry in the database. Variability among NMR structure χ angles is given in Supplemental Table 1.

Assembly of the control *anti* database

To determine the statistical significance of *syn* nucleobase structural features, a control database of *anti* conformation bases was assembled. We used the same RNA molecules as in the *syn*

database and chose 258 *anti* nucleobases using a random number generator as follows: one *anti* nucleobase per structure with zero or one *syn* nucleobases, and two *anti* nucleobases per structure with two or more *syn* nucleobases. As a result, the control database approximated the *syn* data set in size.

Statistical analysis of the database

To assess the significance of *syn* nucleobase data as compared to the *anti* control, the Z-score test was used. The Z-score was defined as:

$$Z = \left| \frac{P_s - P_a}{\sigma} \right| \quad (1)$$

$$\sigma = \sqrt{X(1-X) \left(\frac{1}{n} + \frac{1}{m} \right)} \quad (2)$$

$$X = \frac{s+a}{n+m} \quad (3)$$

where P_s and P_a are the fraction of *syn* and *anti* nucleobases displaying a certain property, respectively (e.g., *syn* nucleobase in binding pocket), s is the number of *syn* nucleobases displaying the property, a is the number of *anti* nucleobases with the property, and n and m represent the total number of *syn* and *anti* nucleobases, respectively. A Z-score >3 is generally considered significant to a confidence level of 99.9% (Anderson and Sclove 1986).

The information content for the 5' and 3' neighbors of *syn* nucleobases was given by the equation (Schneider et al. 1986):

$$R_{\text{Sequence}}(L) = H_{\text{anti}} - H_{\text{syn}}(L) \quad (4)$$

where H is the Shannon uncertainty in bits and calculated by the equation:

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad (5)$$

where P_i is the probability of a particular base appearing as a neighbor to a *syn* nucleobase. To account for the bias of many RNA structures in the PDB toward GC content, the *anti* database was used as a control akin to H_{genome} ; in one instance, this led to a slightly negative value for R_{sequence} (Supplemental Fig. S1).

Pymol was used for all molecular images (DeLano 2002). Kaleidagraph (Synergy Software) was used for plotting of R_{sequence} .

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Yu Zhang, Andrey Krasilnikov, Rick Russell, Doug Turner, Anna Pyle, and Kevin Keating for helpful conversations; and Joshua Boyer for comments on the manuscript. This work

was supported by NSF grant MCB-0527102 to P.C.B. and NIH grant 1F31NS054492-01 to J.E.S.

Received March 29, 2011; accepted July 13, 2011.

REFERENCES

- Abraham M, Dror O, Nussinov R, Wolfson HJ. 2008. Analysis and classification of RNA tertiary structures. *RNA* **14**: 2274–2289.
- Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. 2004. Crystal structure of a self-splicing group I intron with both exons. *Nature* **430**: 45–50.
- Anderson TW, Sclove SL. 1986. *The statistical analysis of data*. Scientific Press, Palo Alto, CA.
- Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, et al. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci* **101**: 6421–6426.
- Batey RT, Gilbert SD, Montange RK. 2004. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* **432**: 411–415.
- Baugh C, Grate D, Wilson C. 2000. 2.8 Å crystal structure of the malachite green aptamer. *J Mol Biol* **301**: 117–128.
- Beard WA, Wilson SH. 2005. Syn-full behavior by T7 DNA polymerase. *Structure* **13**: 1580–1582.
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* **80**: 319–324.
- Bevilacqua PC. 2003. Mechanistic considerations for general acid-base catalysis by RNA: Revisiting the mechanism of the hairpin ribozyme. *Biochemistry* **42**: 2259–2265.
- Bloomfield VA, Crothers DM, Tinoco I Jr. 2000. *Nucleic acids: Structures, properties, and functions*. University Science Books, Sausalito, CA.
- Briebe LG, Kokoska RJ, Bebenek K, Kunkel TA, Ellenberger T. 2005. A lysine residue in the fingers subdomain of T7 DNA polymerase modulates the miscoding potential of 8-oxo-7,8-dihydroguanosine. *Structure* **13**: 1653–1659.
- Chadalavada DM, Senchak SE, Bevilacqua PC. 2002. The folding pathway of the genomic hepatitis delta virus ribozyme is dominated by slow folding of the pseudoknots. *J Mol Biol* **317**: 559–575.
- Chen JH, Yajima R, Chadalavada DM, Chase E, Bevilacqua PC, Golden BL. 2010. A 1.9 Å crystal structure of the HDV ribozyme pre-cleavage suggests both Lewis acid and general acid mechanisms contribute to phosphodiester bond cleavage. *Biochemistry* **49**: 6508–6518.
- Cheong C, Varani G, Tinoco I Jr. 1990. Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature* **346**: 680–682.
- Cochrane JC, Lipchock SV, Strobel SA. 2007. Structural investigation of the GlnS ribozyme bound to its catalytic cofactor. *Chem Biol* **14**: 97–105.
- David L, Lambert D, Gendron P, Major F. 2001. Leadzyme. *Methods Enzymol* **341**: 518–540.
- Davis IW, Murray LW, Richardson JS, Richardson DC. 2004. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res* **32**: W615–W619.
- Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB III, Snoeyink J, Richardson JS, et al. 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**: W375–W383.
- DeLano WL. 2002. *The PyMOL molecular graphics system*. DeLano Scientific, San Carlos, CA.
- Dias E, Battiste JL, Williamson JR. 1994. Chemical probe for glycosidic conformation in telomeric DNAs. *J Am Chem Soc* **116**: 4479–4480.
- Dieckmann T, Suzuki E, Nakamura GK, Feigon J. 1996. Solution structure of an ATP-binding RNA aptamer reveals a novel fold. *RNA* **2**: 628–640.
- Duarte CM, Pyle AM. 1998. Stepping through an RNA structure: A novel approach to conformational analysis. *J Mol Biol* **284**: 1465–1478.
- Fedor MJ. 2009. Comparative enzymology and structural biology of RNA self-cleavage. *Annu Rev Biophys* **38**: 271–299.
- Fox GE. 2010. Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol* **2**: a003483. doi: 10.1101/cshperspect.a003483.
- Garst AD, Heroux A, Rambo RP, Batey RT. 2008. Crystal structure of the lysine riboswitch regulatory mRNA element. *J Biol Chem* **283**: 22347–22351.
- Gautheret D, Major F, Cedergren R. 1993. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J Mol Biol* **229**: 1049–1064.
- Gendron P, Lemieux S, Major F. 2001. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol* **308**: 919–936.
- Golden BL, Kim H, Chase E. 2005. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat Struct Mol Biol* **12**: 82–89.
- Guo F, Gooding AR, Cech TR. 2004. Structure of the *Tetrahymena* ribozyme: Base triple sandwich and metal ion at the active site. *Mol Cell* **16**: 351–362.
- Hammond NB, Tolbert BS, Kierzek R, Turner DH, Kennedy SD. 2010. RNA internal loops with tandem AG pairs: the structure of the 5'GAGU/3'UGAG loop can be dramatically different from others, including 5'AAGU/3'UGAA. *Biochemistry* **49**: 5817–5827.
- Hemmes PR, Oppenheimer L, Jordan F. 1974. Ultrasonic relaxation evaluation of the thermodynamics of syn-anti glycosidic isomerization in adenosine. *J Am Chem Soc* **96**: 6023–6026.
- Hoogstraten CG, Legault P, Pardi A. 1998. NMR solution structure of the lead-dependent ribozyme: Evidence for dynamics in RNA catalysis. *J Mol Biol* **284**: 337–350.
- Keating KS, Pyle AM. 2010. Semiautomated model building for RNA crystallography using a directed rotameric approach. *Proc Natl Acad Sci* **107**: 8177–8182.
- Kitayner M, Rozenberg H, Rohs R, Suad O, Rabinovich D, Honig B, Shakked Z. 2010. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat Struct Mol Biol* **17**: 423–429.
- Klein DJ, Ferre-D'Amare AR. 2006. Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science* **313**: 1752–1756.
- Korostelev A, Ermolenko DN, Noller HF. 2008. Structural dynamics of the ribosome. *Curr Opin Chem Biol* **12**: 674–683.
- Lemieux S, Chartrand P, Cedergren R, Major F. 1998. Modeling active RNA structures using the intersection of conformational space: Application to the lead-activated ribozyme. *RNA* **4**: 739–749.
- Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16**: 279–287.
- Macelvey C, Salter JD, Krucinska J, Wedekind JE. 2008. Structural effects of nucleobase variations at key active site residue Ade38 in the hairpin ribozyme. *RNA* **14**: 1600–1616.
- Major F. 2003. Building three-dimensional ribonucleic acid structures. *Comput Sci Eng* **5**: 44–53.
- Martick M, Scott WG. 2006. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **126**: 309–320.
- Maruta S, Ohki T, Kambara T, Ikebe M. 1998. Characterization of the interaction of myosin with ATP analogues having the *syn* conformation with respect to the adenine-ribose bond. *Eur J Biochem* **256**: 229–237.
- McCown PJ, Roth A, Breaker RR. 2011. An expanded collection and refined consensus model of glmS ribozymes. *RNA* **17**: 728–736.
- Murthy VL, Srinivasan R, Draper DE, Rose GD. 1999. A complete conformational map for RNA. *J Mol Biol* **291**: 313–327.
- Nikolova EN, Kim E, Wise AA, O'Brien PJ, Andricioaei I, Al-Hashimi HM. 2011. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470**: 498–502.
- Noller HF, Hoffarth V, Zimniak L. 1992. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256**: 1416–1419.

- Osborne SE, Ellington AD. 1997. Nucleic acid selection and the challenge of combinatorial chemistry. *Chem Rev* **97**: 349–370.
- Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**: 51–55.
- Proctor DJ, Kierzek E, Kierzek R, Bevilacqua PC. 2003. Restricting the conformational heterogeneity of RNA by specific incorporation of 8-bromoguanosine. *J Am Chem Soc* **125**: 2390–2391.
- Proctor DJ, Ma H, Kierzek E, Kierzek R, Gruebele M, Bevilacqua PC. 2004. Folding thermodynamics and kinetics of YNMG RNA hairpins: Specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. *Biochemistry* **43**: 14004–14014.
- Reiter NJ, Osterman A, Torres-Larios A, Swinger KK, Pan T, Mondragon A. 2010. Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature* **468**: 784–789.
- Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, HersHKovits E, Williams LD, et al. 2008. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14**: 465–481.
- Rupert PB, Massey AP, Sigurdsson ST, Ferre-D'Amare AR. 2002. Transition state stabilization by a catalytic RNA. *Science* **298**: 1421–1424.
- Saenger W. 1984. *Principles of nucleic acid structure*. Springer-Verlag, New York.
- Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**: 415–431.
- Schwalbe CH, Saenger W. 1973. 6-Azaauridine, a nucleoside with unusual ribose conformation. The molecular and crystal structure. *J Mol Biol* **75**: 129–143.
- Sclavi B, Sullivan M, Chance MR, Brenowitz M, Woodson SA. 1998. RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science* **279**: 1940–1943.
- Shi H, Moore PB. 2000. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* **6**: 1091–1105.
- Smith TF, Lee JC, Gutell RR, Hartman H. 2008. The origin and evolution of the ribosome. *Biol Direct* **3**: 16. doi:10.1186/1745-6150-3-16.
- Smith KD, Lipchock SV, Livingston AL, Shanahan CA, Strobel SA. 2010. Structural and biochemical determinants of ligand binding by the c-di-GMP riboswitch. *Biochemistry* **49**: 7351–7359.
- Stahley MR, Strobel SA. 2005. Structural evidence for a two-metal-ion mechanism of group I intron splicing. *Science* **309**: 1587–1590.
- Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem Sci* **28**: 411–418.
- Tanner NK, Schaff S, Thill G, Petit-Koskas E, Crain-Denoyelle AM, Westhof E. 1994. A three-dimensional model of hepatitis delta virus ribozyme based on biochemical and mutational analyses. *Curr Biol* **4**: 488–498.
- Toor N, Keating KS, Taylor SD, Pyle AM. 2008. Crystal structure of a self-spliced group II intron. *Science* **320**: 77–82.
- Tucker BJ, Breaker RR. 2005. Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* **15**: 342–348.
- Veeraraghavan N, Ganguly A, Chen JH, Bevilacqua PC, Hammes-Schiffer S, Golden BL. 2011. Metal binding motif in the active site of the HDV ribozyme binds divalent and monovalent ions. *Biochemistry* **50**: 2672–2682.
- Webb C-HT, Riccitelli NJ, Ruminiski DJ, Lupták A. 2009. Widespread occurrence of self-cleaving ribozymes. *Science* **326**: 953. doi:10.1126/science.1178084.
- Wedekind JE, McKay DB. 2003. Crystal structure of the leadzyme at 1.8 Å resolution: Metal ion binding and the implications for catalytic mechanism and allo site ion regulation. *Biochemistry* **42**: 9554–9563.
- Wilson TJ, Lilley DM. 2011. Do the hairpin and VS ribozymes share a common catalytic mechanism based on general acid-base catalysis? A critical assessment of available experimental data. *RNA* **17**: 213–221.
- Xin Y, Laing C, Leontis NB, Schlick T. 2008. Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA* **14**: 2465–2477.
- Yajima R, Proctor DJ, Kierzek R, Kierzek E, Bevilacqua PC. 2007. A conformationally restricted guanosine analog reveals the catalytic relevance of three structures of an RNA enzyme. *Chem Biol* **14**: 23–30.
- Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M. 1996. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science* **272**: 1343–1347.
- Yang H, Zhan Y, Fenn D, Chi LM, Lam SL. 2008. Effect of 1-methyladenine on double-helical DNA structures. *FEBS Lett* **582**: 1629–1633.
- Zimmermann GR, Wick CL, Shields TP, Jenison RD, Pardi A. 2000. Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. *RNA* **6**: 659–667.

RNA

A PUBLICATION OF THE RNA SOCIETY

Prevalence of *syn* nucleobases in the active sites of functional RNAs

Joshua E. Sokoloski, Stephanie A. Godfrey, Sarah E. Dombrowski, et al.

RNA 2011 17: 1775-1787 originally published online August 26, 2011

Access the most recent version at doi:[10.1261/rna.2759911](https://doi.org/10.1261/rna.2759911)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2011/08/01/rna.2759911.DC1>

References

This article cites 70 articles, 22 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/17/10/1775.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>